Accepted Manuscript

A balanced scheduler with data reuse and replication for scientific workflows in cloud computing systems

Israel Casas, Javid Taheri, Rajiv Ranjan, Lizhe Wang, Albert Y. Zomaya

PII: S0167-739X(15)00388-X
DOI: http://dx.doi.org/10.1016/j.future.2015.12.005
Reference: FUTURE 2918

To appear in: *Future Generation Computer Systems*

Received date: 13 April 2015
Revised date: 13 October 2015
Accepted date: 5 December 2015

Please cite this article as: I. Casas, J. Taheri, R. Ranjan, L. Wang, A.Y. Zomaya, A balanced scheduler with data reuse and replication for scientific workflows in cloud computing systems, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2015.12.005

A Balanced Scheduler with Data Reuse and Replication for Scientific Workflows in Cloud Computing Systems

Israel Casas [a,d], Javid Taheri [a], Rajiv Ranjan [b], Lizhe Wang [c], Albert Y. Zomaya [a]
[a] School of Information Technologies, The University of Sydney Sydney, Australia
[b] School of Computing Science, Newcastle University, Newcastle-Upon-Tyne, United Kingdom
[d] Data61, CSIRO, Australia
[c] School of Computer Science, China University of Geosciences, Wuhan 430074, China.

ARTICLE INFO

Article history:
Received
Received in revised form
Accepted
Available online

Keywords:
Cloud computing
Scientific Workflow
Scheduling
Virtual Machine
Data-intensive computing
Big data

ABSTRACT

Cloud computing provides substantial opportunities to researchers who demand pay-as-you-go computing systems. Although cloud provider (e.g., Amazon Web Services) and application provider (e.g., biologists, physicists, and online gaming companies) both have specific performance requirements (e.g. application response time), it is the cloud scheduler's responsibility to map the application to underlying cloud resources. This article presents a Balanced and file Reuse-Replication Scheduling (BaRRS) algorithm for cloud computing environments to optimally schedule scientific application workflows. BaRRS splits scientific workflows into multiple sub-workflows to balance system utilization via parallelization. It also exploits data reuse and replication techniques to optimize the amount of data that needs to be transferred among tasks at run-time. BaRRS analyzes the key application features (e.g., task execution times, dependency patterns and file sizes) of scientific workflows for adapting existing data reuse and replication techniques to cloud systems. Further, BaRRS performs a trade-off analysis to select the optimal solution based on two optimization constraints: execution time and monetary cost of running scientific workflows. BaRRS is compared with a state-of-the-art scheduling approach; experiments prove its superior performance. Experiments include four well known scientific workflows with different dependency patterns and data file sizes. Results were promising and also highlighted most critical factors affecting execution of scientific applications on clouds.

1. Introduction

Cloud computing has great impact on Information Technology (IT) solutions for both scientific and business applications [42, 47]. Cloud computing environment has important features that are important for both science and business applications/purposes. Clouds also offer solutions to computationally intensive applications similar to HPC (High Performance Computing) environments such as supercomputing centers [43]. From the business perspective, clouds offer flexible platforms to both cloud providers and application owners. Cloud computing offers an unique computing ecosystem where providers and application owners can establish elastic relationship driven by application performance requirements (e.g. availability, execution time, monetary budget, etc.) and characteristics (e.g. input data size, number of end-users connecting to that application, output data size, etc.) [44, 48].

Current computing applications demand research on cloud environments for their efficient use of resources [45]. Cloud providers are concerned and need to maintain their services in a relatively unreliable environment, while producing profitable revenues. Cloud application owners, on the other hand, want services/resources that meet strict performance requirements. Both require access to a cloud scheduler framework and algorithm that can automatically map applications to cloud resources while ensuring application-level performance and provider-level resource utilization goals [46, 49].

Due to their data intensive nature, modern scientific applications can benefit if cloud schedulers include data reuse and replication techniques in executing their workflows [36, 37, 38]. Data reuse avoids file transfers by